D524–D529 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 30 November 2012*
doi:10.1093/nar/gks1169

PDBTM: Protein Data Bank of transmembrane proteins after 8 years

Dániel Kozma[1], István Simon[2] and Gábor E. Tusnády[1,*]

[1]Lendület Membrane Protein Bioinformatics Research Group and [2]Protein Structure Research Group, Institute of Enzymology, MTA RCNS, PO Box 7, H-1518 Budapest, Hungary

Received September 21, 2012; Revised October 25, 2012; Accepted October 28, 2012

ABSTRACT

The PDBTM database (available at http://pdbtm.enzim.hu), the first comprehensive and up-to-date transmembrane protein selection of the Protein Data Bank, was launched in 2004. The database was created and has been continuously updated by the TMDET algorithm that is able to distinguish between transmembrane and non-transmembrane proteins using their 3D atomic coordinates only. The TMDET algorithm can locate the spatial positions of transmembrane proteins in lipid bilayer as well. During the last 8 years not only the size of the PDBTM database has been steadily growing from ~400 to 1700 entries but also new structural elements have been identified, in addition to the well-known α-helical bundle and β-barrel structures. Numerous 'exotic' transmembrane protein structures have been solved since the first release, which has made it necessary to define these new structural elements, such as membrane loops or interfacial helices in the database. This article reports the new features of the PDBTM database that have been added since its first release, and our current efforts to keep the database up-to-date and easy to use so that it may continue to serve as a fundamental resource for the scientific community.

INTRODUCTION

Transmembrane proteins play an important role in the living cells for energy production, regulation and metabolism. The fact that half of present-day drugs have some effect on transmembrane proteins (1,2) also underlines their biological importance. Furthermore, ~25% of the human genome might code transmembrane proteins (3), which means about 5–6000 structures. Due to the structural and physiochemical properties of these proteins, the experimental techniques for structure determination are not so straightforward. As a consequence, the proportion of transmembrane and globular proteins in the Protein Data Bank (PDB) (4) database is <2% according to the PDBTM database (5,6). Hence, the PDBTM database was created in 2004 to collect these cases. The PDBTM database was the first to address the problems of transmembrane protein structures in the PDB database, namely the fact that these proteins cannot be identified using the annotation in the PDB's entries. Therefore, a new method was needed, which is based on only the 3D coordinates to identify transmembrane segments and does not require additional information. Moreover, since one of the most important environments, the double lipid layer, is not part of the solved atomic structures due to the experimental difficulties of structure determination, theoretical methods are required to determine the orientations of the transmembrane proteins relative to the lipid bilayer. We developed a method, called TMDET (7), which addresses and solves the above-mentioned problems. Since then several transmembrane databases have become available on the Internet, utilizing different theoretical algorithms and techniques, and serving different purposes. For the sake of comparability, let us briefly summarize the main properties of such databases.

The OPM (8) contains a well-structured classification of membrane proteins. The orientation of the protein relative to the membrane normal is defined by minimizing its transfer energy ($\Delta G_{transfer}$) from water to the lipid bilayer with respect to the shift along the bilayer normal, hydrophobic thickness, rotation angle and tilt angle (9). Some missing side-chain atoms are added and the structure of residues at the water–lipid interface is adjusted. The results of these calculations are used to transform the atomic coordinates of integral membrane proteins in a way that the membrane normal be parallel with the *z*-axis. In the OPM database, the transformed coordinate files contain membrane planes too, which are represented by dummy oxygen and nitrogen atoms. The topology data about transmembrane proteins are also given in the OPM database, i.e. what part of the proteins face to the cytosolic space and what part to the extra-cytosolic one.

*To whom correspondence should be addressed. Tel: +36 1 279 3159; Fax: +36 1 279 3108; Email: tusnady.gabor@ttk.mta.hu

The CGDB (10) database contains the final system coordinates of coarse-grained simulation-relaxed transmembrane protein structures in bilayer and their analysis from the aspect of protein–lipid interaction. This database has the most sophisticated model in terms of physics, as it utilizes a previously developed high-throughput computational approach to perform the coarse-grained simulations. There are two other analogous databases which are more specific: the KDB is for K-channels (http://sbcb.bioch.ox.ac.uk/kdb/) and the OMPDB is a set of outer membrane proteins obtained by full-atom simulations (11). These databases contain indispensable information on dynamic aspects and stability.

One of the most reliable database of membrane proteins is the membrane proteins of known structure (Mpstruct, http://blanco.biomol.uci.edu/membrane_proteins_xtal.html), which is regularly updated. In this, membrane proteins are classified using a simpler classification scheme than the one used by the OPM. Although the OPM and the PDBTM contain information about the membrane orientation of proteins and about the classification of sequence segments, the Mpstruct does not.

There are several other databases collecting transmembrane proteins and some of their properties (12–16): (i) the MPDB (12) is a relational database of structural and functional information on integral, anchored and peripheral membrane proteins and peptides derived from the literature and from the PDB database. It provides various search parameters (protein characteristics, structure determination methods, crystallization techniques, detergents, temperature, 'pH', authors, etc.) and records are linked to the PDB, the Pfam (13) or the PubMed. It is a weekly updated database following the PDB weekly updates. In addition, the MPDB provides different statistics about the sources and the detergents used in crystallization, as well as about applied expression systems, among other data. (ii) The TMFunction (14) is a collection of >2900 experimentally observed functional residues in membrane proteins. Each entry in the TMFunction database includes the numerical values for the parameters IC50, V(max), relative activity of mutants with respect to wild-type protein, binding affinity and dissociation constant. (iii) The Transporter Classification Database (15) is a web accessible, curated, relational database containing sequence, classification, structural, functional and evolutionary information about transport systems from a variety of living organisms.

In the PDBTM database, we collect all transmembrane proteins for which structures have been solved so far; we check and if necessary correct their biologically active oligomer form given in PDB files, define their membrane orientation and set their transmembrane segments, membrane re-entrant loops and interfacial helices (IFHs).

NEW FEATURES OF THE PDBTM DATABASE

Although the main architecture of the TMDET algorithm has not been changed, several extensions have been added to the basic algorithm to enhance the usability and reliability of our database. The need for the new features is the consequence of the development this scientific field has experienced. We have enhanced the database to include those structural elements, which were not known or were rarely represented when the database was created. These are IFHs and re-entrant regions (loop, hairpin and re-entrant coil) (17). These and some other new features will be discussed in the following sections.

Correcting biomatrices

The biological form of the protein usually does not correspond to the molecule, which is present in the asymmetric unit. Therefore, the symmetry operations, which need to be applied to generate the active oligomer form, are displayed in the PDB file in the BIOMOLECULE section as a matrix transformation, called biomatrix. The oligomer form usually is defined by the authors or is calculated by theoretical calculations using PQS (18) or PISA (19). Both of these algorithms have been developed to determine the quaternary structure of globular proteins, therefore they may fail when applied to transmembrane proteins. We have found several files, where the crystals contain the biologically active oligomer form, but the BIOMOLECULE records are set improperly (e.g. 2atk, 2jk5, 2zld) and those, where the crystals contain oligomer forms that do not exist in the membrane. These latter cases cannot be recognized by the above-mentioned methods. Most frequently they are subunits with anti-parallel orientation in a homo-dimer transmembrane protein, which were discussed in our original article (5). The usage of inappropriate biomatrices occasionally leads to the inaccurate definitions of the orientation of membrane proteins relative to the membrane. In some cases, it could be a $\sim 20^{\circ}$ or a larger difference between monomer and oligomer forms.

We aimed to identify and correct problems, which can be associated with biomatrices and leads to incorrect oligomers. Therefore, we developed a new algorithm, which uses homologous protein structures to generate biomatrices for proteins with inappropriate biomatrix in the PDB. The outline of the protocol is as follows. Protein structures having only one chain without any biomatrix annotation (or only the identity matrix is given in the biomatrix records) are selected in one pool, whereas those which have only one chain and a biomatrix were stored in an other pool. Then a BLAST search is performed against the sequences of the second pool for each sequence of the first one. The protein with the highest hit is used as a candidate and if the sequential similarity is >90%, then the query structure will be superimposed on the candidate using TM-align (20) algorithm. TM-align gives the transformation ($\widehat{\mathbf{T}}$), which turns $\mathbf{P}_{\mathbf{query}}$ to $\mathbf{P}_{\mathbf{target}}$ formally:

$$\widehat{\mathbf{T}}\mathbf{P}_{\mathbf{query}} = \mathbf{P}_{\mathbf{target}}. \quad (1)$$

Assuming that there are $\mathbf{P}_{\mathbf{query}}$ and $\mathbf{P}_{\mathbf{target}}$ identical monomer structures with different absolute coordinates and the corresponding biomatrices are $\widehat{\mathbf{B}}_{\mathbf{query}}$ and $\widehat{\mathbf{B}}_{\mathbf{target}}$, then we get:

$$\widehat{\mathbf{T}}\widehat{\mathbf{B}}_{\mathbf{query}}\mathbf{P}_{\mathbf{query}} = \widehat{\mathbf{B}}_{\mathbf{target}}\mathbf{P}_{\mathbf{target}}. \quad (2)$$

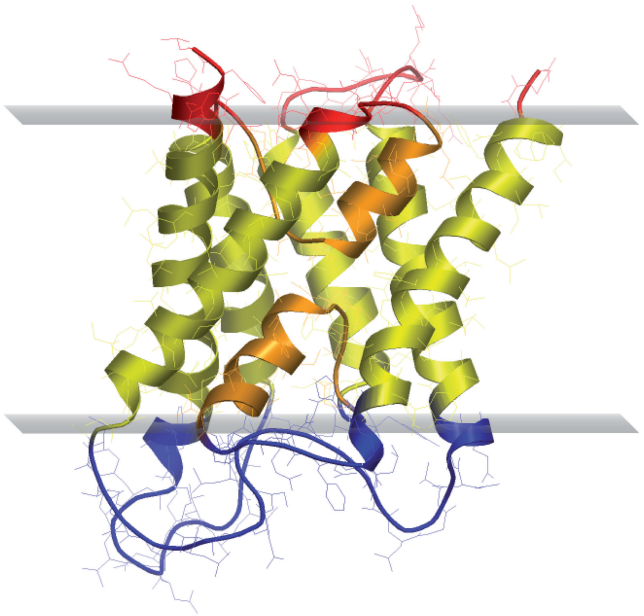

Figure 1. Loops (coloured in orange) in 1h6i, a refined structure of human aquaporin (22).

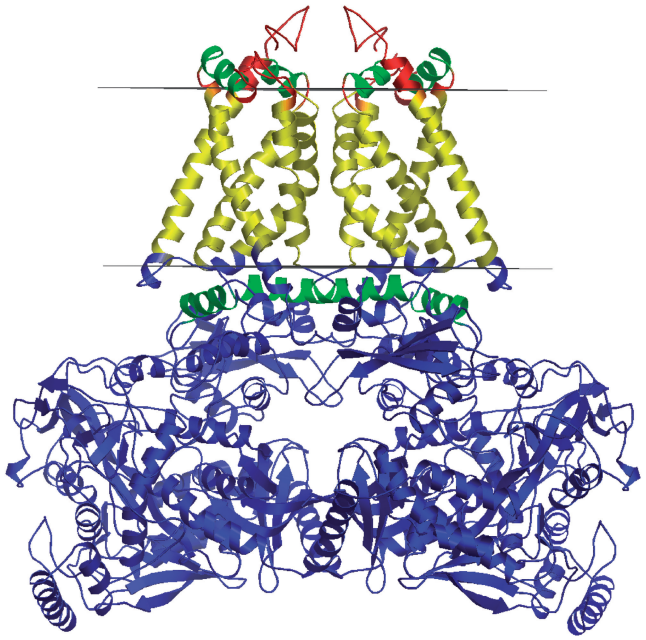

Figure 2. IFH (coloured in green) in 1e7p, a quinol-fumarate reductase from *Wolinella succinogenes* (28).

Replacing $\mathbf{P}_{\mathbf{target}}$ with $\widehat{\mathbf{T}}\mathbf{P}_{\mathbf{query}}$ on the bases of Equation (1), in Equation (2), we obtain:

$$\widehat{\mathbf{T}}\widehat{\mathbf{B}}_{\mathbf{query}}\mathbf{P}_{\mathbf{query}} = \widehat{\mathbf{B}}_{\mathbf{target}}\widehat{\mathbf{T}}\mathbf{P}_{\mathbf{query}}. \quad (3)$$

Hence

$$\widehat{\mathbf{T}}\widehat{\mathbf{B}}_{\mathbf{query}} = \widehat{\mathbf{B}}_{\mathbf{target}}\widehat{\mathbf{T}}, \quad (4)$$

$$\widehat{\mathbf{B}}_{\mathbf{query}} = \widehat{\mathbf{T}}^{-1}\widehat{\mathbf{B}}_{\mathbf{target}}\widehat{\mathbf{T}}. \quad (5)$$

We have checked the accuracy of this procedure by applying it on those entries, which are homo-oligomer molecules and have correct BIOMOLECULE record. The PDBTM database contains 318 such entries. After sequence filtering to 90% identity, we got 57 entries. We could generate biomatrices for 43 entries using homologous protein structures. After calculating the coordinates using these newly generated biomatrices, we calculate the root mean square deviation (RMSD) between the original and computed coordinates. The RMSD values of 40 out of the 43 entries were <1 Å (avg: 0.38 ± 0.20 Å), while the worst alignment produced a 3.3 Å RMSD.

In cases, when the crystal contains the correct oligomer form, but this is not given in the BIOMOLECULE record, we supply the correct crystallographic symmetry transformation. Altogether, the biomatrices of 34 entries have been corrected. The largest tilt angle difference between the corrected and uncorrected original forms was found in the case of 2w0f, a potassium-channel KcsA–Fab complex with tetraoctylammonium. In the PDB file, it appears as a monomer (after applying the given biomatrix transformation), but its active form is tetramer. The angle deviation was 23° and the region borders moved up to four residues. We have found similar angle deviation in the OPM database as well. The largest tilt angle deviation, 19° in the OPM database, can be found between 1py6 and 1m0l. 1py6 is a monomeric protein in the PDB, while 1m0l is a homo-trimer of the same bacteriorhodopsin.

Membrane re-entrant loops

Membrane re-entrant loops with both ends facing the same side of the membrane were first detected in the late 90 s (21) in the case of the cardiac Na^+/Ca^{2+} exchanger. Later it was shown that several other channel-like transmembrane proteins contain this type of structural element, e.g. aquaporins (22), potassium channels (23), chloride channels (24), etc. (Figure 1). We have developed a new algorithm as an extension of the TMDET to detect these structural elements using only the 3D atomic coordinates of given transmembrane proteins and the transformation matrices produced by the TMDET algorithm, by searching sequence segments having both end on the same side of the membrane, and diving into the membrane with at least 6 Å (measured from the membrane–water interface). This algorithm can detect any type of re-entrant loops (e.g. helix–loop–coil, coil–loop–helix, coil–loop–coil), but the database currently does not contain these pieces of information. Currently, there are 258 proteins in the PDBTM database, which contain one or more re-entrant loops.

Interfacial helices

Another newly implemented structural class is IFHs that are α-helices laying in the membrane–water interface parallel to the membrane plane (Figure 2). They have various structural roles, for example, they are responsible for the regulation of channel gating in both the KirBac 1.1 inward rectifying potassium channel (25) and the MscS mechanosensitive channel (26), while in photosystem I,

IFHs appear to shield cofactors from the aqueous phase (27).

A further extension of the TMDET algorithm contains a subroutine which identifies these regions. First, we collect α-helical regions not in the membrane, and longer than four residues, and calculate the tilt angle relative to the membrane plane and the distance from the membrane–water boundary. The algorithm uses two threshold parameters: the distance (<9 Å) from the membrane–water boundary and the tilt angle (<30°). As a result of this extension, we have identified IFHs in 851 proteins.

THE NEW USER INTERFACE OF THE PDBTM

The homepage of the upgraded version of the PDBTM database utilizes the Wt C++ Web Toolkit (http://www.webtoolkit.eu/wt) programming library and the OpenAstexViewer (29) to visualize transmembrane protein structures highlighted with different colours for the different region types to make the structure even more informative. We have recently created a complex web application for investigating protein 3D structures and residue–residue interactions (30), where both the Wt and the OpenAstexViewer have been successfully utilized.

The PDBTM entry viewer

The layout of the PDBTM molecule viewer can be seen in Figure 3. The navigation bar (Figure 3A) contains an up-to-date list of IDs of current transmembrane protein structures in the PDBTM database. The arrows serve for the navigation in this list. The previous structure viewer has been replaced with the OpenAstexViewer (29). The colouring of the 3D structure (Figure 3B) and sequence (Figure 3C) is identical in order to help users to find sequence segments more easily in the 3D structure.

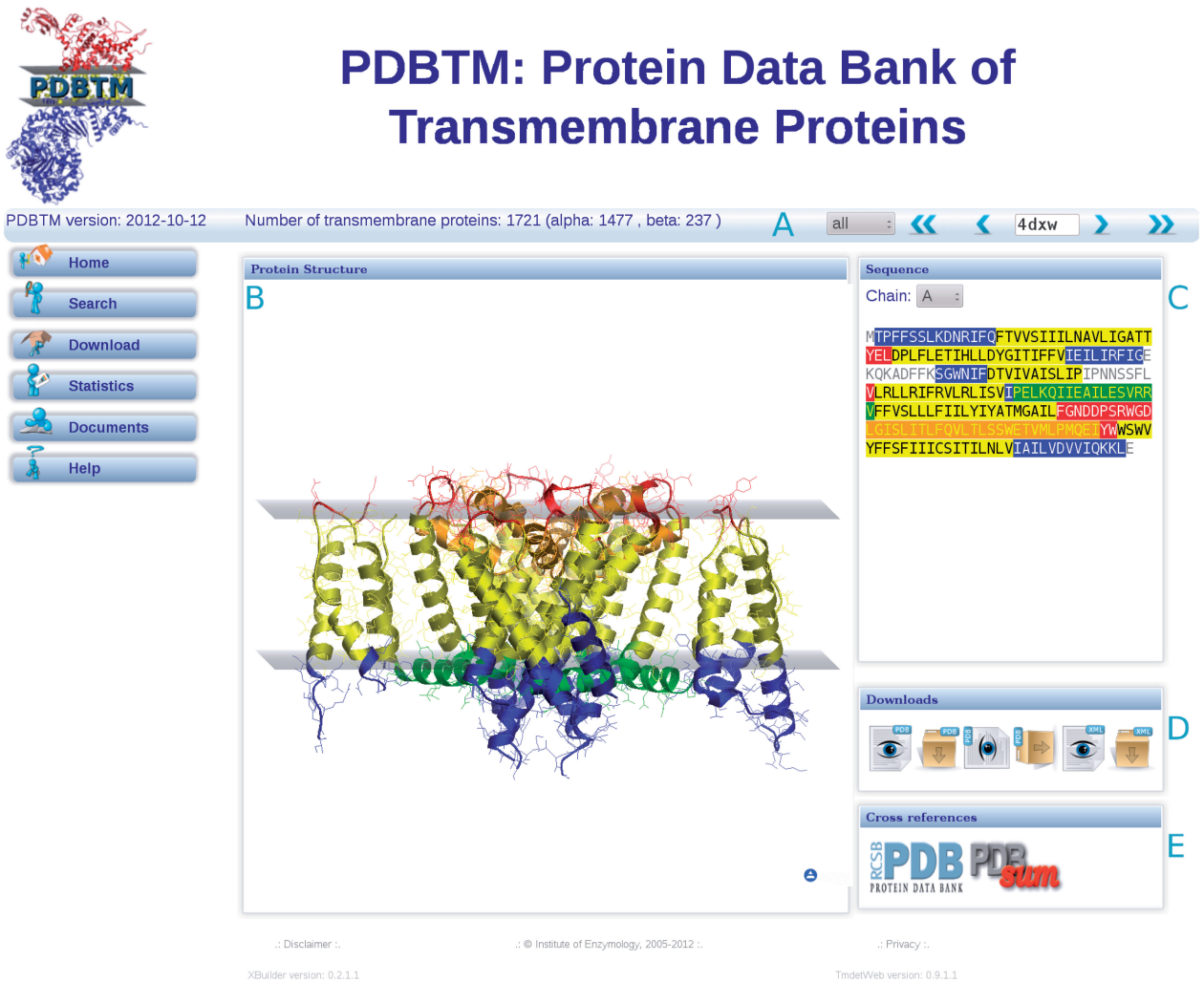

Figure 3. The PDBTM entry viewer. (**A**) The navigation bar which is always visible for the sake of comfortable and instant navigation. Using the arrows one can navigate to the first entry, step back, step forward or jump to the end. (**B**) The structure viewer (29), using the same colours as in the sequence box. (**C**) Sequence box, containing the chain selector and the sequence of the actual protein chain. (**D**) File download section, where the user can download or simply view the original and the transformed PDB files as well as PDBTM XML files. (**E**) Cross-reference links to the RCSB PDB and PDBsum (31) databases.

These two widgets are connected through signals, so by clicking on any sequence regions (except the grey-coloured ones, which represent sequence without solved structure), the representation of the corresponding residues in the structure viewer turns from cartoon to sphere.

Users can download or simply view the original and the transformed PDB files as well as the PDBTM XML files (Figure 3D), which describe the regions of the structure, chain sequences and all the necessary information to build up the transformed PDB structure from the original one.

Advanced search system

The web server allows users to perform various types of search in the database. Some ordinary, frequently used search requests have already been implemented, but users can also query custom requests, either in a form field or by using the address line of the browser. This latest feature enables the users to refer to their query results as a constantly updated list by bookmarking the given query. The search results can be browsed or downloaded as a whole in various file formats. For more detailed description visit the manual of the PDBTM (http://pdbtm.enzim.hu/?_=/help/manual).

CONCLUSION

The PDBTM database is a comprehensive, up-to-date and continuously updated transmembrane protein database. As of today, it contains >1700 entries whose regions are classified into structural elements such as transmembrane helices, transmembrane beta segments, membrane re-entrant loops or IFHs. The flexible search method makes data mining easier for bioinformaticians who are interested in transmembrane proteins and their structures. All kinds of feedback and advice are most welcome, as they will help us to improve and to satisfy the diverse demands of users more fully.

ACKNOWLEDGEMENTS

Comments on the article by Mónika Fuxreiter and László Benke and on the manual of the PDBTM database by Bálint Mészáros are gratefully acknowledged. We would like to express our gratitude for the help of Koen Deforche and István Reményi in the development of PDBTM.

FUNDING

Hungarian Scientific Research Fund (OTKA) [NK100482 and K104586]; 'Lendület' Program of the Hungarian Academy of Sciences (to G.E.T.). Funding for open access charge: 'Lendület' Program of the Hungarian Academy of Sciences.

Conflict of interest statement. None declared.

REFERENCES

1. Overington,J.P., Al-Lazikani,B. and Hopkins,A.L. (2006) How many drug targets are there? *Nat. Rev. Drug Discov.*, **5**, 993–996.
2. Parrill,A.L. (2008) Crystal structures of a second G protein-coupled receptor: triumphs and implications. *ChemMedChem*, **3**, 1021–1023.
3. Fagerberg,L., Jonasson,K., von Heijne,G., Uhlén,M. and Berglund,L. (2010) Prediction of the human membrane proteome. *Proteomics*, **10**, 1141–1149.
4. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlic,A., Quesada,M., Quinn,G.B., Westbrook,J.D. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.
5. Tusnády,G.E., Dosztányi,Z. and Simon,I. (2004) Transmembrane proteins in the Protein Data Bank: identification and classification. *Bioinformatics*, **20**, 2964–2972.
6. Tusnády,G.E., Dosztányi,Z. and Simon,I. (2005) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res.*, **33**, D275–D278.
7. Tusnády,G.E., Dosztányi,Z. and Simon,I. (2005) TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics*, **21**, 1276–1277.
8. Lomize,M.A., Pogozheva,I.D., Joo,H., Mosberg,H.I. and Lomize,A.L. (2012) OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res.*, **40**, D370–D376.
9. Lomize,A.L., Pogozheva,I.D., Lomize,M.A. and Mosberg,H.I. (2006) Positioning of proteins in membranes: a computational approach. *Protein Sci.*, **15**, 1318–1333.
10. Chetwynd,A.P., Scott,K.A., Mokrab,Y. and Sansom,M.S.P. (2008) CGDB: a database of membrane protein/lipid interactions by coarse-grained molecular dynamics simulations. *Mol. Membr. Biol.*, **25**, 662–669.
11. Tsirigos,K.D., Bagos,P.G. and Hamodrakas,S.J. (2011) OMPdb: a database of beta-barrel outer membrane proteins from Gram-negative bacteria. *Nucleic Acids Res.*, **39**, D324–D331.
12. Raman,P., Cherezov,V. and Caffrey,M. (2006) The Membrane Protein Data Bank. *Cell. Mol. Life Sci.*, **63**, 36–51.
13. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
14. Gromiha,M.M., Yabuki,Y., Suresh,M.X., Thangakani,A.M., Suwa,M. and Fukui,K. (2009) TMFunction: database for functional residues in membrane proteins. *Nucleic Acids Res.*, **37**, D201–D204.
15. Saier,M.H., Ming,R.Y., Keith,N., Dorjee,G.T. and Charles,E. (2009) The Transporter Classification Database: recent advances. *Nucleic Acids Res.*, **37**, D274–D278.
16. Gromiha,M.M., Yabuki,Y., Kundu,S., Suharnan,S. and Suwa,M. (2007) TMBETA-GENOME: database for annotated beta-barrel membrane proteins in genomic sequences. *Nucleic Acids Res.*, **35**, D314–D316.
17. Nugent,T. and Jones,D.T. (2011) Membrane protein structural bioinformatics. *J. Struct. Biol.*, **179**, 327–337.
18. Henrick,K. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358–361.
19. Krissinel,E. and Henrick,K. (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.*, **372**, 774–797.
20. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
21. Iwamoto,T., Nakamura,T.Y., Pan,Y., Uehara,A., Imanaga,I. and Shigekawa,M. (1999) Unique topology of the internal repeats in the cardiac Na^+/Ca^{2+} exchanger. *FEBS Lett.*, **446**, 264–268.
22. de Groot,B.L., Engel,A. and Grubmüller,H. (2001) A refined structure of human aquaporin-1. *FEBS Lett.*, **504**, 206–211.
23. Zhou,Y., Morais-Cabral,J.H., Kaufman,A. and MacKinnon,R. (2001) Chemistry of ion coordination and hydration revealed by a K^+ channel–Fab complex at 2.0 Å resolution. *Nature*, **414**, 43–48.

24. Dutzler,R., Campbell,E.B., Cadene,M., Chait,B.T. and MacKinnon,R. (2002) X-ray structure of a ClC chloride channel at 3.0 Å reveals the molecular basis of anion selectivity. *Nature*, **415**, 287–294.
25. Doyle,D.A. (2004) Structural themes in ion channels. *Eur. Biophys. J.*, **33**, 175–179.
26. Bass,R.B., Locher,K.P., Borths,E., Poon,Y., Strop,P., Lee,A. and Rees,D.C. (2003) The structures of BtuCD and MscS and their implications for transporter and channel function. *FEBS Lett.*, **555**, 111–115.
27. Jordan,P., Fromme,P., Witt,H.T., Klukas,O., Saenger,W. and Krauss,N. (2001) Three-dimensional structure of cyanobacterial photosystem I at 2.5 Å resolution. *Nature*, **411**, 909–917.
28. Lancaster,C.R., Gross,R. and Simon,J. (2001) A third crystal form of Wolinella succinogenes quinol:fumarate reductase reveals domain closure at the site of fumarate reduction. *Eur. J. Biochem.*, **268**, 1820–1827.
29. Hartshorn,M.J. (2002) AstexViewer: a visualisation aid for structure-based drug design. *J. Comput. Aided Mol. Des.*, **16**, 871–881.
30. Kozma,D., Simon,I. and Tusnády,G.E. (2012) CMWeb: an interactive on-line tool for analysing residue–residue contacts and contact prediction methods. *Nucleic Acids Res.*, **40**, W329–W333.
31. Laskowski,R.A. (2009) PDBsum new things. *Nucleic Acids Res.*, **37**, D355–D359.